OIPE
MAR 3 1 2005
PATENT & TRADEMARK OFFICE
JC112

Attorney Docket No. YOR920000430US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Patent Application

I hereby certify that this paper is being deposited on this date with the U.S. Postal Service as first class mail addressed to the Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

Signature: Lori L. Culpis Date: March 28, 2005

Applicant(s): C. Aggarwal et al.
Docket No.: YOR920000430US1
Serial No.: 09/703,174
Filing Date: October 31, 2000
Group: 2176
Examiner: Nathan Hillery

Title: Methods and Apparatus for Intelligent Crawling on the World Wide Web

---

APPEAL BRIEF

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

Applicants hereby appeal the final rejection dated September 23, 2004 of claims 1-27 of the above-identified application.

REAL PARTY IN INTEREST

The present application is assigned to International Business Machines Corp., as evidenced by an assignment recorded October 31, 2000 in the U.S. Patent and Trademark Office at Reel 11290, Frame 0654. The assignee, International Business Machines Corp., is the real party in interest.

RELATED APPEALS AND INTERFERENCES

There are no known related appeals or interferences.

04/01/2005 MAHMED1 00000001 500510 09703174
01 FC:1402 500.00 DA

STATUS OF CLAIMS

The present application was filed on October 31, 2000 with claims 1-27. Claims 1-27 are currently pending in the application. Claims 1, 10 and 19 are the independent claims.

Each of claims 1-27 stands finally rejected under 35 U.S.C. §103(a). Claims 1-27 are appealed.

STATUS OF AMENDMENTS

There have been no amendments filed subsequent to the final rejection.

SUMMARY OF CLAIMED SUBJECT MATTER

Independent claim 1 is directed to a computer-based method of performing document retrieval in accordance with an information network. The method comprises the steps of retrieving one or more documents from the information network that satisfy a user-defined predicate, collecting statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed, and using the collected statistical information to automatically determine further document retrieval operations. Independent claims 10 and 19 recite similar limitations.

As illustratively explained in the present application, the present invention provides methods and apparatus for performing intelligent crawling. Particularly, the intelligent crawling techniques of the invention provide a crawler mechanism which is capable of learning as it crawls in order to focus the search for documents on the information network being explored, e.g., world wide web. This crawler mechanism stores information about the crawled documents as it retrieves the documents, and then uses the information to further focus its search appropriately. The inventive techniques result in the crawling of a small percentage of the documents on the world wide web (Specification, page 3, lines 2-9).

The present invention provides a more interesting and significantly more general alternative to conventional crawling techniques. As is evident from the inventive teachings, no specific model for web linkage structure is assumed in intelligent crawling according to the invention. Rather, the crawler gradually learns the linkage structure statistically as it progresses. By linkage structure, this

refers to the fact that there is a certain relationship between the content of a web page and the candidates that it links to. For example, a web page containing the word "Edmund Guide" is likely to link to web pages on automobile dealers. In general, linkage structure refers to the relationship between the various features of a web page such as content, tokens in Universal Resource Locators (URL), etc. Further, in general, it is preferred that the linkage structure be predicate-dependent. An intelligent crawler according to the invention learns about the linking structure during the crawl and find the most relevant pages. Initially, the crawler behavior may be as random as a general crawler but it then gradually starts auto-focusing as it encounters documents which satisfy the predicate. A certain level of supervision in terms of documents which satisfy the predicate may be preferred since it would be very helpful in speeding up the process (especially for very specific predicates), but is not essential for the framework of the invention. This predicate may be a decision predicate or a quantitative predicate which assigns a certain level of priority to the search (Specification, page 4, line 22, through page 5, line 13).

The intelligent crawler of the invention may preferably be implemented as a graph search algorithm which works by treating web pages as nodes and links as edges. The crawler keeps track of the nodes which it has already visited, and for each node, it decides the priority in which it visits based on its understanding of which nodes is likely to satisfy the predicate. Thus, at each point the crawler maintains candidate nodes which it is likely to crawl and keeps re-adjusting the priority of these nodes as its information about linkage structure increases (Specification, page 5, lines 14-20).

FIG. 2 of the present application is a flow diagram illustrating a process for intelligent web crawling according to an embodiment of the invention. This illustrative process may be implemented by the computer system 10 shown in FIG. 1. The illustrative embodiment of the intelligent crawling method shown in FIG. 2 recursively crawls the URL pages and maintains two sets of statistical information: (1) aggregate statistical information; and (2) predicate-specific statistical information. These two sets of information may be used in order to effectively find the web pages which are most related to a given predicate. The aggregate statistical information is maintained on all the retrieved web pages. This aggregate information is updated every time a web page is retrieved. The predicate-specific information is updated every time a web page belonging

to a given predicate is retrieved. The aggregate information is likely to be different from the predicate-specific information. This difference is used to determine which of the URL documents are more likely than others to belong to the predicate. The input to the method of FIG. 2 is a list of URLs from which the crawl starts. In addition, a predicate which is used to focus and control the crawl is input. These inputs may be provided by a user (Specification, page 8, lines 10-25).

GROUND OF REJECTION TO BE REVIEWED ON APPEAL

Claims 1-27 are rejected under 35 U.S.C. §103(a) as being unpatentable over U.S. Patent No. 6,112,203 to Bharat et al. (hereinafter "Bharat").

ARGUMENT

Regarding the §103(a) rejection of claims 1-27, Appellants respectfully submit that, as is well-established law, an obviousness rejection under §103(a) requires that the cited reference must "teach or suggest all the claim limitations," and that there be some suggestion or motivation, either in the reference itself or in the knowledge generally available to one of ordinary skill in the art, to modify the reference teachings. See Manual of Patent Examining Procedure (MPEP), Eighth Edition, August 2001, §706.02(j).

Appellants respectfully submit that the Bharat reference fails to teach or suggest all the claim limitations that the final Office Action asserts it does, and that there is no cogent motivation for modifying the reference teachings, as asserted by the final Office Action, to reach the claimed invention.

The present invention, for example, as recited in independent claim 1, provides a computer-based method of performing document retrieval in accordance with an information network. The method comprises the steps of retrieving one or more documents from the information network that satisfy a user-defined predicate, collecting statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed, and using the collected statistical information to automatically determine further document retrieval operations. Independent claims 10 and 19 recite similar limitations.

As illustratively explained in the present specification at page 3, lines 2-9:

> The present invention provides methods and apparatus for performing intelligent crawling. Particularly, the intelligent crawling techniques of the invention provide a crawler mechanism which is capable of learning as it crawls in order to focus the search for documents on the information network being explored, e.g., world wide web. This crawler mechanism stores information about the crawled documents as it retrieves the documents, and then uses the information to further focus its search appropriately. The inventive techniques result in the crawling of a small percentage of the documents on the world wide web. (Underlining added for emphasis)

In contrast, as pointed out in Appellants' first response dated May 4, 2004 and in their response to final dated December 23, 2004, Bharat discloses a method for ranking documents in a hyperlinked environment using connectivity and content analysis (see Abstract of Bharat). That is, Bharat does not disclose an intelligent crawling technique that is able to further focus its search appropriately.

More particularly, Bharat does not disclose the step of "collecting statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed," as in the claimed invention. While Bharat does disclose content analysis, it does not appear that any "statistical information" is being collected in the Bharat document ranking technique.

Further, the final Office Action again expressly acknowledges, at page 4, that Bharat does not disclose the step of "using the collected statistical information to automatically determine further document retrieval operations," as in the claimed invention. However, the final Office Action again suggests that it would have been obvious to modify Bharat to include such a step.

Appellants again submit that it is the step of "using the collected statistical information to automatically determine further document retrieval operations," that even more particularly distinguishes the document ranking technique of Bharat from the technique of the claimed invention. That is, Bharat merely prunes a "start set" in order to determine a best ranking to be presented to a user. As explained in Bharat, the start set is a set of data objects derived from a response to a search engine query, see, e.g., column 4, lines 49-52. That is, once the start set is retrieved, Bharat merely processes the retrieved set of objects so as to rank the retrieved objects in a manner that will be most helpful to the user.

However, because Bharat has nothing to do with an intelligent crawling technique, it does not teach or suggest the step of "using the collected statistical information to automatically determine further document retrieval operations," as in the claimed invention. Bharat merely uses its analysis to rank the already-retrieved results and not to automatically determine further document retrieval operations.

In the "Response to Arguments" section of the final Office Action, the Examiner points out that the phrase "an intelligent crawling technique that is able to further focus its search appropriately" is not recited in the claims. Appellants realize this and offered the illustrative explanatory language as a way of giving the Examiner a better understanding of the distinguishing features of the invention that are expressly claimed. For example, the intelligent crawling technique of the invention may be able to further focus its search appropriately by "using the collected statistical information to automatically determine further document retrieval operations." Thus, it is the document retrieval method (e.g., intelligent crawling technique) that uses the collected statistical information to automatically determine further document retrieval operations. It seems that the Examiner does not appreciate this distinction since, at page 3 of the final Office Action, it is stated that "[t]he skilled artisan would be motivated to use the collected statistical information . . . ." It is not the user that uses this collected statistical information, but rather the document retrieval method (e.g., intelligent crawling technique) that uses the collected statistical information to automatically determine further document retrieval operations. In fact, Bharat does not teach any step of determining further document retrieval operations, whether by a user or automatically.

Furthermore, despite the Examiner's contention, there is nothing in Bharat that would provide motivation to modify the document ranking techniques of Bharat to yield an intelligent crawling technique of the invention. More particularly, Appellants assert that there is no motivation to modify Bharat to use collected statistical information to automatically determine further document retrieval operations since Bharat only deals with ranking already-retrieved documents.

Furthermore, as pointed out in Appellants previous responses, the Federal Circuit has stated that when patentability turns on the question of obviousness, the obviousness determination "must be based on objective evidence of record" and that "this precedent has been reinforced in myriad

decisions, and cannot be dispensed with." In re Sang-Su Lee, 277 F.3d 1338, 1343 (Fed. Cir. 2002). Moreover, the Federal Circuit has stated that "conclusory statements" by an examiner fail to adequately address the factual question of motivation, which is material to patentability and cannot be resolved "on subjective belief and unknown authority." Id. at 1343-1344.

In the final Office Action at page 3, the Examiner provides the following statements to prove motivation to modify Bharat, with emphasis supplied: "[t]he skilled artisan would be motivated to use the collected statistical information based on this teaching because the invention relates generally to . . . ranking retrieved documents based on content . . . and because a good ranking process will return 'useful' pages."

Appellants again submit that these statements are based on the type of "subjective belief and unknown authority" that the Federal Circuit has indicated provides insufficient support for an obviousness rejection. More specifically, other than citing two sentences in Bharat about the merits of ranking retrieved documents (col. 1, lines 7-9; col. 4, lines 20-21), the Examiner fails to identify any objective evidence of record which supports the proposed modification to Bharat. Again, the final Office Action suggests that the process of ranking already-retrieved documents provides motivation for the conclusion that Bharat could be modified to use collected statistical information to automatically determine further document retrieval operations. However, this is clearly not a reasonable conclusion based at least on the fact that Bharat is completely silent as to any determination of further document retrieval operations.

For at least the above reasons, Appellants respectfully assert that independent claims 1, 10 and 19 are patentable over Bharat.

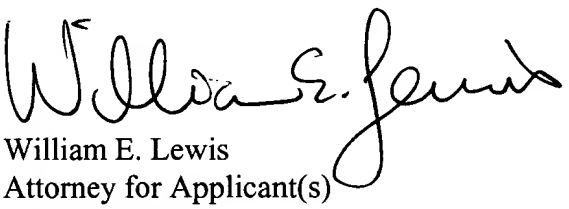
The remainder of the claims (namely, claims 2-9, 11-18 and 20-27) rejected over Bharat depend, either directly or indirectly, from claims 1, 10 or 19, which are believed patentable for the reasons set forth above. Furthermore, the remaining claims define additional patentable subject matter in their own right.

By way of example only, claims 5, 14 and 23 recite wherein the collected statistical information is used to direct further document retrieval operations toward documents which are more likely to satisfy the predicate than would otherwise occur with respect to document retrieval

operations that are not directed using the collected statistical information. Further, claims 6, 15 and 24 recite wherein the collected statistical information is used to direct further document retrieval operations toward documents which are similar to the one or more retrieved documents that also satisfy the predicate. Still further, claims 7, 16 and 25 recite wherein the collected statistical information is used to direct further document retrieval operations toward documents which are linked to by other documents which also satisfy the predicate. Since Bharat does not teach, suggest, or provide motivation to be modified to provide a system that is able to direct further document retrieval operations, such claims are clearly patentable over Bharat.

In view of the above, Appellants believe that claims 1-27 are in condition for allowance, and respectfully requests the withdrawal of the §103(a) rejections.

Respectfully submitted,



Date: March 28, 2005

William E. Lewis
Attorney for Applicant(s)
Reg. No. 39,274
Ryan, Mason & Lewis, LLP
90 Forest Avenue
Locust Valley, NY 11560
(516) 759-2946

CLAIMS APPENDIX

1. A computer-based method of performing document retrieval in accordance with an information network, the method comprising the steps of:

retrieving one or more documents from the information network that satisfy a user-defined predicate;

collecting statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed; and

using the collected statistical information to automatically determine further document retrieval operations.

2. The method of claim 1, wherein the user-defined predicate specifies content associated with a document.

3. The method of claim 1, wherein the statistical information collection step uses content of the one or more retrieved documents.

4. The method of claim 1, wherein the statistical information collection step considers whether the user-defined predicate has been satisfied by the one or more retrieved documents.

5. The method of claim 1, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are more likely to satisfy the predicate than would otherwise occur with respect to document retrieval operations that are not directed using the collected statistical information.

6. The method of claim 1, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are similar to the one or more retrieved documents that also satisfy the predicate.

7. The method of claim 1, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are linked to by other documents which also satisfy the predicate.

8. The method of claim 1, wherein the information network is the world wide web and a document is a web page.

9. The method of claim 8, wherein the statistical information collection step uses one or more uniform resource locator tokens in the one or more retrieved web pages.

10. Apparatus for performing document retrieval in accordance with an information network, the apparatus comprising:

at least one processor operative to: (i) retrieve one or more documents from the information network that satisfy a user-defined predicate; (ii) collect statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed; and (iii) use the collected statistical information to automatically determine further document retrieval operations.

11. The apparatus of claim 10, wherein the user-defined predicate specifies content associated with a document.

12. The apparatus of claim 10, wherein the statistical information collection operation uses content of the one or more retrieved documents.

13. The apparatus of claim 10, wherein the statistical information collection operation considers whether the user-defined predicate has been satisfied by the one or more retrieved documents.

14. The apparatus of claim 10, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are more likely to satisfy the predicate than would otherwise occur with respect to document retrieval operations that are not directed using the collected statistical information.

15. The apparatus of claim 10, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are similar to the one or more retrieved documents that also satisfy the predicate.

16. The apparatus of claim 10, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are linked to by other documents which also satisfy the predicate.

17. The apparatus of claim 10, wherein the information network is the world wide web and a document is a web page.

18. The apparatus of claim 17, wherein the statistical information collection operation uses one or more uniform resource locator tokens in the one or more retrieved web pages.

19. An article of manufacture for performing document retrieval in accordance with an information network, comprising a machine readable medium containing one or more programs which when executed implement the steps of:

retrieving one or more documents from the information network that satisfy a user-defined predicate;

collecting statistical information about the one or more retrieved documents as the one or more retrieved documents are analyzed; and

using the collected statistical information to automatically determine further document retrieval operations.

20. The article of claim 19, wherein the user-defined predicate specifies content associated with a document.

21. The article of claim 19, wherein the statistical information collection step uses content of the one or more retrieved documents.

22. The article of claim 19, wherein the statistical information collection step considers whether the user-defined predicate has been satisfied by the one or more retrieved documents.

23. The article of claim 19, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are more likely to satisfy the predicate than would otherwise occur with respect to document retrieval operations that are not directed using the collected statistical information.

24. The article of claim 19, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are similar to the one or more retrieved documents that also satisfy the predicate.

25. The article of claim 19, wherein the collected statistical information is used to direct further document retrieval operations toward documents which are linked to by other documents which also satisfy the predicate.

26. The article of claim 19, wherein the information network is the world wide web and a document is a web page.

27. The article of claim 26, wherein the statistical information collection step uses one or more uniform resource locator tokens in the one or more retrieved web pages.

EVIDENCE APPENDIX

None

RELATED PROCEEDINGS APPENDIX

None